Course Title: **Information Systems Design** Course Code: CCE4235 4th year
Date: 13.6.2016 (Second term) Allowed time: 3 hrs

Answer the following questions:

Question No. 1 **(15 marks)**

For each of the following, please circle the letter introducing the best answer- each one is worth one mark:

1. Which word or phrase completes the statement? A spreadsheet to a data island is as a centralized database to a _____?
   a) Data Warehouse
   b) Data Repository
   c) Analytic Sandbox
   d) Data Mart
2. You are studying the behavior of a population, and you are provided with multidimensional data at the individual level. You have identified four specific individuals who are valuable to your study, and would like to find all users who are most similar to each individual. Which algorithm is the most appropriate for this study?
   a) Linear regression
   b) Association rules
   c) K-means clustering
   d) Decision trees
3. In which lifecycle stage the analytic sandbox is prepared?
   a) Discovery
   b) Model planning
   c) Model building
   d) Data preparation
4. When would you use a Wilcoxson Rank Sum test?
   a) When the data can easily be sorted
   b) When you cannot make an assumption about the distribution of the populations
   c) When the populations represent the sums of other values
   d) When the data cannot easily be sorted
5. A data scientist wants to predict the probability of death from heart disease based on three risk factors: age, gender, and blood cholesterol level. What is the most appropriate method for this project?
   a) Linear regression
   b) Logistic regression
   c) K-means clustering
   d) Apriori algorithm
6. Consider the example of an analysis for fraud detection on credit card usage. You will need to ensure higher risk transactions that may indicate fraudulent credit card activity are retained in your data for analysis, and not dropped as outliers during pre-processing. What will be your approach for loading data into the analytical sandbox for this analysis?
   a) ETL
   b) EDW
   c) ELT
   d) OLTP
7. In which lifecycle stage are initial hypotheses formed?
   a) Discovery
   b) Model planning
   c) Model building
   d) Data preparation

8. A disk drive manufacturer has a defect rate of less than 2% with 98% confidence. A quality assurance team samples 1000 disk drives and finds 14 defective units. Which action should the team recommend?
   a) The manufacturing process should be inspected for problems.
   b) A larger sample size should be taken to determine if the plant is functioning properly
   c) A smaller sample size should be taken to determine if the plant is functioning properly
   d) The manufacturing process is functioning properly and no further action is required.
9. Which characteristic applies only to Business Intelligence as opposed to Data Science?
   a) Supports solving "what if" scenarios
   b) Uses large data sets
   c) Uses only structured data
   d) Uses predictive modeling techniques
10. Which activity might be performed in the Operationalize phase of the Data Analytics Lifecycle?
   a) Try different analytical techniques
   b) Try different variables
   c) Transform existing variables
   d) Run a pilot
11. You are asked to create a model to predict the total number of monthly subscribers for a specific magazine. You are provided with one - years' worth of subscription and payment data, user demographic data, and 10-years' worth of content of the magazine (articles and pictures). Which algorithm is the most appropriate for building a predictive model for subscribers?
   a) Linear regression
   b) Logistic regression
   c) Decision trees
   d) TF-IDF
12. Your organization has a website where visitors randomly receive one of two coupons. It is also possible that visitors to the website will not receive a coupon. You have been asked to determine if offering a coupon to visitors to your website has any impact on their purchase decision. Which analysis method should you use?
   a) K-means clustering
   b) Association rules
   c) Student T-test
   d) One-way ANOVA
13. When would you prefer a Naive Bayes model to a logistic regression model for classification?
   a) When you need to estimate the probability of an outcome not just which class it is in
   b) When all the input variables are numerical.
   c) When you are using several categorical input variables with over 1000 possible values each.
   d) When some of the input variables might be correlated.
14. Which data asset is an example of quasi-structured data?
   a) Web clickstream data
   b) XML data file
   c) Database table
   d) D. News article
15. What is an example of a null hypothesis?
   a) that a newly created model does not provide better predictions than the currently existing model
   b) that a newly created model provides a prediction of a null sample mean
   c) that a newly created model provides a prediction of a null population mean
   d) that a newly created model provides a prediction that will be well fit to the null distribution

Question (2)

1. students were given different drug treatments before revising for their exams. Some were given a memory drug, some placebo drug and some no treatment. The exam scores(%) are given below for three different groups. Carry out one-way ANOVA to test the hypothesis that the treatments will have different effects.

	Memory drug	Placebo	No treatment
	70	37	3
	77	43	10
	83	50	17
	90	57	23
	97	63	30
Mean	83.40	50.00	16.60
variance	112.30	109.00	112.30
Grand mean		50.00	
Grand variance		892.14	

steps: the mean

$$m_1 = \frac{70 + 77 + 83 + 90 + 97}{5} = 83.4$$

$$m_2 = \frac{37 + 43 + 50 + 57 + 63}{5} = 50$$

$$m_3 = \frac{3 + 10 + 17 + 23 + 30}{5} = 16.6$$

$$m_0 = \frac{m_1 + m_2 + m_3}{3} = \frac{83.4 + 50 + 16.6}{3} = 50$$

step 2: sum of squares

$$SS_{within} = \sum_J (X_1^J - m_1)^2 + \sum_J (X_2^J - m_2)^2 + \sum_J (X_3^J - m_3)^2$$

$$\hookrightarrow (70 - 83.4)^2 + (77 - 83.4)^2 + (82 - 83.4)^2 + (90 - 83.4)^2 + (97 - 83.4)^2$$

$$SS_{within} = 449.2 + 436 + 449.2 = 1334.4$$

$$SS_{total} = \sum_i \sum_J (X_i^J - m_0) = 12490$$

$$SS_{Between} = SS_{total} - SS_{within} = 11155.6$$

step 3: F

N → no. of items
K → no. of attributes

$$S_B^2 = \frac{SS_{Between}}{K-1} = \frac{11155.6}{3-1} = 5577.8$$

$$S_W^2 = \frac{SS_{within}}{N-K} = \frac{1334.4}{15-3} = 111.2$$

$$F = \frac{S_B^2}{S_W^2} = 50.1$$

$F > 1$

reject Null Hypothesis.

[2] Consider the set of items is I = {milk, bread, butter, beer} and small database of transactions containing the items (where 1 codes presence & 0 codes absence of item in transaction) shown in table.

a) Apply Apriori algorithm (let minimum support is 40 %) to find all frequent item set in database.

b) use these frequent item sets and the minimum confidence constraint (let minimum support = 70 %) to form association rules.

Transaction ID	milk	Bread	Butter	beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0
6	1	0	0	0
7	0	1	1	1
8	1	1	1	1
9	0	1	0	1
10	1	1	0	0
11	1	0	0	0
12	0	0	0	0
13	1	1	1	0
14	1	0	1	0
15	1	1	1	1

* 15 transaction ⇒ 40% = 6

step 1

C_1

itemset
milk
bread
Butter
beer

⇒

L_1

Itemset	sup.
milk	9
bread	10
butter	7
beer	5 — X

step 2

C_2

Item set
milk, bread
milk, butter
bread, butter

⇒

L_2

Itemset	sup.
milk, bread	6
milk, butter	5 — X
bread, butter	6

step 3

C_3

item set
milk, bread, butter

L_3

itemset	sup
milk, bread, butter	4 — X

* Candidate rules from L_2

1) milk → bread

2) bread → milk

3) bread → butter

4) butter → bread

Rule	set	cnt	set	cnt	Confidence
if milk then bread	milk	9	milk & bread	6	$\frac{6}{9} = 66.6\%$
if bread then milk	bread	10	bread & milk	6	$\frac{6}{10} = 60\%$
if bread then butter	bread	10	bread & butter	6	$\frac{6}{10} = 60\%$
if butter then bread	butter	7	bread & butter	6	$\frac{6}{7} = 85.7\%$

* we want Confidence $\geq 70\%$

rule is : if butter then bread.

Q2.3) Explain the difference between

Business intelligence	Data science
↳ structured data ↳ traditional sources ↳ manageable datasets	↳ structured / unstructured. ↳ multiple types of sources. ↳ very large datasets.
standard	optimization, Predictive modeling, statistical analysis
(His questions) ↳ How many did we sell ? ↳ where is the problem ?	(His questions) ↳ what if ..? * open-ended questions

Question 3

1. Consider this training dataset

X_1	X_2	X_3	X_4
1	1	1	0
1	1	0	0
0	0	0	0
0	1	0	1
1	0	1	1
0	1	1	1

Apply the Naïve Bayesian classifier to this data set and compute the probability score for $P(y=1|X)$ for $X=(1,0,0)$

$X_1=1$ & $X_2=0$ & $X_3=0$

$$P(y=1|X) = \frac{P(X|y=1)\,P(y=1)}{P(X)}$$

مش متأكد من الحل ده

$$= \frac{\frac{1}{3} * \frac{1}{3} * \frac{1}{3} * \frac{3}{6}}{\frac{3}{6} * \frac{2}{6} * \frac{2}{6}} = \frac{2}{9} = 0.22$$

2. How ROC Curve is used to diagnose the effectiveness of logistic regression model?

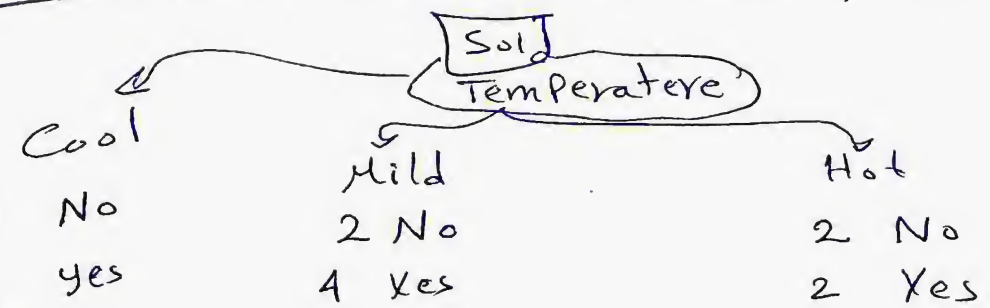
Thanks for un-signed heroes.

Q3.2 How Roc Curve is used to diagnose the effectiveness of logistic regression model?

↳ answered in Lec 10 Page 12,13

Q3.3 Consider the following dataset... Compute the information gain of the temperature attribute.

outlook	temperature	humidity	windy	class
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no



$$H(t) = -\left(\frac{5}{14}\log\frac{5}{14} + \frac{9}{14}\log\frac{9}{14}\right) = 0.94$$

No Yes

$$H_{attr} = \frac{4}{14}\left(\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}\right) +$$

$$\frac{6}{14}\left(\frac{2}{6}\log\frac{2}{6} + \frac{4}{6}\log\frac{4}{6}\right) +$$

$$\frac{4}{14}\left(\frac{2}{4}\log\frac{2}{4} + \frac{2}{4}\log\frac{2}{4}\right) = \boxed{0.911}$$

$$\text{Info-gain} = 0.94 - 0.911 = 0.029$$

$$H = -\sum_c P(c)\log(P(c))$$

$$H_{attr} = -\sum_v P(v)\sum_c P(c|v)\log P(c|v)$$

$$\text{Info-gain} = H - H_{attr}$$

لو فيه سؤال عن ال confusion matrix بس مش حاليا.